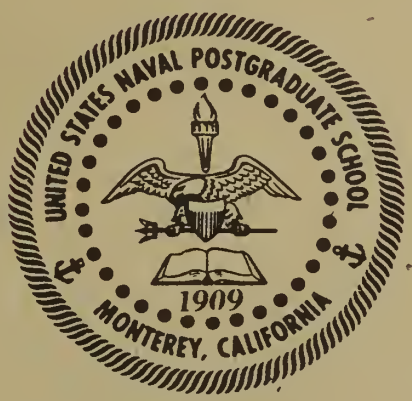
Michael Satosi Watanabe

A STUDY OF ERGODICITY AND REDUNDANCY BASED ON INTER-SYMBOL CORRELATION OF FINITE RANGE.

UNITED STATES
NAVAL POSTGRADUATE SCHOOL



A STUDY OF ERGODICITY AND REDUNDANCY BASED ON INTERSYMBOL CORRELATION OF FINITE RANGE

— By —

Michael SATOSI WATANABE,
Professor of Physics

Research Paper No. 4

A STUDY OF
ERGODICITY AND REDUNDANCY
BASED ON
INTERSYMBOL CORRELATION
OF FINITE RANGE

By

SATOSI WATANABE

A Study of Ergodicity and Redundancy based on Intersymbol Correlation of Finite Range

by

Satosi Watanabe

United States Naval Postgraduate School
Monterey, California

Some of the basic concepts of information theory are critically reviewed in the light of a generalized formulation of the theory of Markoff's chains, in which the initial and final states are sequences of symbols of different lengths, and occurrence of symbols is governed by intersymbol correlation of finite range. In particular, the conditions of ergodicity and the structure of "ergodic subsets" of sequences of arbitrary length are carefully discussed. A mathematical method is developed to determine the "range" and "strength" of intersymbol correlation. A brief summary of the content is given at the end of Section 1.

#1. Introduction

The aim of this paper is to clarify some of the basic, but often carelessly used concepts of information theory, viz., the concepts of ergodicity, intersymbol correlation and redundancy. There are two approaches to this problem-complex pertaining to probability. One is an empirical point of view, and probability here is understood in its statistical aspect. The other is an a priori point of view which deals with probability mainly in its predictive aspect. In the first standpoint, the entire population of messages in a language is supposed to be given, and the various probabilities are calculated by the actual frequencies of individual symbols or those of sequences of symbols. According to this method, a unique value of the probability of appearance of a given symbol or a given sequence can be statistically determined. In the second point of view, an ensemble of messages is supposed to be engendered by the given correlation probabilities starting from a given initial symbol or a given initial sequence of symbols. In this case, the existence of a unique, non-vanishing value of the probability of appearance of a given symbol or a given sequence is not guaranteed, for it may vanish with increasing length of messages, and it may depend on the initial condition. Thus, the problem of ergodicity acquires foremost importance in this approach.

Our section 2 dealing with the problem of ergodicity is therefore developed in the framework of the second point of view. Once the nature of the ergodicity condition is clarified and this condition is

assumed to be fulfilled, then a smooth passage from the second point of view to the first becomes easy. Thus, our section 3 on redundancy can be interpreted in either point of view.

It is not implied by the foregoing paragraphs that the problem of ergodicity is irrelevant to the first standpoint or cannot be formulated in the framework of this standpoint. The situation is that the nucleus of the problem under consideration can be exhibited more directly and naturally in the second point of view.

The usual theory of Markoff's chains, which is based on transition probabilities from one state to another, is extended in this paper to the case where the probability $Q(a_1, \ldots, a_{\nu-1} \mid a_\nu)$ of symbol a_ν appearing in a message is dependent on the $(\nu - 1)$ immediately preceding symbols, ν being the range of intersymbol correlation. A population of infinitely long messages is considered to be engendered solely by this intersymbol correlation probability: $Q(a_1, \ldots, a_{\nu-1} \mid a_\nu)$ from a given $(\nu - 1)$-symbol initial sequence. The problem of ergodicity then pertains to existence of unique (i.e., independent of initial sequence), non-vanishing value of $P(a_1, \ldots, a_\mu)$, which should give the probability that a μ-symbol sequence arbitrarily taken from the population is $(a_1, \ldots, a_\mu)$, μ being not necessarily equal to ν. This generalized problem of ergodicity is discussed in our Section 2.

It is shown not only that finiteness of correlation range does not warrant ergodicity, as is often erroneously assumed in existing literature, but also that if $\mu < \nu$ the quantity P can have more

than one finite value depending on the initial sequence, a situation which does not exist in the ordinary Markoff chains.

Under the conditions that guarantee existence of unique (whether or not non-vanishing) value of P, a convenient quantity, called correlation index W_μ, defined by Eq. (31), is introduced, characterizing both "range" and "strength" of correlation. First, it represents the "range", in the sense that the actual correlation range is the maximum value of μ for which $W_\mu \neq 0$. This criterion is both of theoretical and practical interest. Theoretically, this determines the applicability of the generalized theory of Markoff's chains, and practically, this can be used to measure the existing correlation range in a given population of messages.

Second, this quantity W_μ represents the "strength" of correlation, in the sense that W_μ quantitatively measures the decrease of information due to the existence of μ - symbol correlation as compared with the $(\mu - 1)$ - symbol correlation. Finally the so-called redundancy is expressed in the form of a compact series in ascending range-numbers of the correlation indices, Eq. (42).

#2. Ergodicity

We assume the alphabet under consideration to consist of N symbols: S_1 , S_2 , . . . S_N . We shall constantly use a mathematical symbol:

$$Q(a_1, a_2, .., a_m | a_{m+1}, .., a_{n-1}, a_n) , \tag{1}$$

where each one of a_1 , a_2 , . ., a_n can be any one of the N symbols.

Definition I. The quantity denoted by (1) represents the probability that the last (n - m) symbols of a sequence of n symbols are (a_{m+1}, . ., a_n) when it is known that the first m symbols of the sequence are (a_1 , . ., a_m).

By the very nature of probability, we have

$$\begin{aligned} &Q(a_1, .., a_m | a_{m+1}, .., a_n) \geq 0 , \\ &\sum_{a_{m+1}} .. \sum_{a_n} Q(a_1, .., a_m | a_{m+1}, .., a_n) = 1 . \end{aligned} \tag{2}$$

If there is no correlation between symbols, the probability of any place in a sequence being occupied by symbol S_i is independent of the preceding symbols. As result, the only quantity which determines a probability of the type (1) is $Q(S_i)$ which represents the probability of symbol S_i appearing at any one place. In this case, we have:

$$\begin{aligned} &Q(a_1, .., a_m | a_{m+1}, .., a_n) \\ &= Q(a_{m+1}) Q(a_{m+2}) .. Q(a_n) . \end{aligned}$$

If the correlation extends, for instance, over three consecutive symbols, and not more than three, then the probability of a place in

a sequence being occupied by symbol S_i will depend on the two symbols directly preceding it, but not on the symbols beyond these two. This means that the quantities $Q(S_i, S_j \mid S_k)$ determine the general probability (1): $Q(a_1, \cdots, a_m \mid a_{m+1}, \cdots, a_n)$

$$= Q(a_{m-1}, a_m \mid a_{m+1})\, Q(a_m, a_{m+1} \mid a_{m+2}) \cdots Q(a_{n-2}, a_{n-1} \mid a_n).$$

In general, we have the following theorem:

Theorem I. If the intersymbol correlation does not extend over more than μ consecutive symbols in a sequence, we can factorize (1) as follows:

$$Q(a_1, \cdots, a_m \mid a_{m+1}, \cdots, a_n)$$

$$= Q(a_{m-\mu+2}, \cdots, a_m \mid a_{m+1})\, Q(a_{m-\mu+3}, \cdots, a_{m+1} \mid a_{m+2}) \cdots Q(a_{n-\mu+1}, \cdots, a_{n-1} \mid a_n) \quad (3)$$

This theorem can be used to define the "range-number" of intersymbol correlation: this number ν is the minimum allowable μ in the decomposition (3).

Assuming the correlation to be of range ν, we consider all the possible sequences whose first $(\nu - 1)$ symbols are given to be, say, $(a_1, a_2, \cdots, a_{\nu-1})$. Among these sequences starting with $(a_1, a_2, \cdots a_{\nu-1})$, we inquire the probability of those sequences whose first ν symbols are $(a_1, b_1, b_2, \cdots, b_{\nu-1})$. This probability is obviously given by

$$R(a_1, a_2, \cdots, a_{\nu-1} \mid b_1, b_2, \cdots, b_{\nu-1}) = Q(a_1, a_2, \cdots, a_{\nu-1} \mid b_{\nu-1})$$

if $(a_2, \cdots, a_{\nu-1}) = (b_1, \cdots, b_{\nu-2})$,

and otherwise

$$R(a_1, a_2, \cdots, a_{\nu-1} \mid b_1, b_2, \cdots, b_{\nu-1}) = 0 \ .$$

In other words, the probability in question can be written in a matrix form:

$$(a_1, a_2, \cdots, a_{\nu-1} \mid R \mid b_1, b_2, \cdots b_{\nu-1})$$

$$= Q(a_1, \cdots, a_{\nu-1} \mid b_{\nu-1})\, \delta(a_2, b_1)\, \delta(a_3, b_2) \cdots \delta(a_{\nu-1}, b_{\nu-2}) \ , \qquad (4)$$

with

$$\delta(S_i, S_j) = 0 \qquad \text{if} \qquad i \neq j$$

$$\delta(S_i, S_j) = 1 \qquad \text{if} \qquad i = j \ .$$

Using this matrix-expression, the probability, in the above population of sequences, of a particular sequence $(b_1, b_2 \ldots b_{\nu-1})$ appearing in such a position that the place distance between a_1 and b_1 is m symbols can be given by

$$T^{(m)}(a_1, \cdots, a_{\nu-1} \mid b_1, \cdots, b_{\nu-1})$$

$$= (a_1, \cdots, a_{\nu-1} \mid R^m \mid b_1, \cdots, b_{\nu-1}) \ , \qquad (5)$$

where R^m simply means the m-th power of R in the sense of matrix-multiplication.

With the help of the quantity (5), we can further calculate the probability of a given sequence of any length $(\mu - 1)$, say $(b_1, \ldots, b_{\mu-1})$, appearing at any position after the initial $(a_1, \ldots, a_{\nu-1})$. If $\mu > \nu$ this probability will be

$$T^{(m)}(a_1, \cdots, a_{\nu-1} \mid b_1, \cdots, b_{\mu-1})$$

$$= T^{(m)}(a_1, \cdots, a_{\nu-1} \mid b_1, \cdots, b_{\nu-1})\, Q(b_1, \cdots, b_{\nu-1} \mid b_\nu) \cdots Q(b_{\mu-\nu}, \cdots, b_{\mu-2} \mid b_{\mu-1}) \qquad (6)$$

where m stands for the symbol distance between a_1 and b_1.

If $\mu<\nu$, we have

$$T^{(m)}(a_1,\cdots,a_{\nu-1}\,|\,b_1,\cdots,b_{\mu-1})$$

$$=\sum_{b_\mu}\cdots\sum_{b_{\nu-1}}T^{(m)}(a_1,\cdots,a_{\nu-1}\,|\,b_1,\cdots,b_{\mu-1},b_\mu,\cdots,b_{\nu-1}), \tag{7}$$

where m bears the same meaning.

Now, the average probability of sequence $(b_1,\cdots,b_{\mu-1})$ with the "place-distance" not larger than m will be

$$U^{(m)}(a_1,\cdots,a_{\nu-1}\,|\,b_1,\cdots,b_{\mu-1})=\frac{1}{m}\sum_{\ell=1}^{m}T^{(\ell)}(a_1,\cdots,a_{\nu-1}\,|\,b_1,\cdots,b_{\mu-1}). \tag{8}$$

We now proceed to define what we mean by ergodicity in this paper. We consider all the possible, infinitely long sequences which start with a given initial sequence $(a_1,\cdots,a_{\nu-1})$ and ask the <u>average</u> probability of the sequence $(b_1,\cdots,b_{\mu-1})$ appearing in any position. This probability evidently has the mathematical expression:

$$\lim_{m\to\infty}U^{(m)}(a_1,\cdots,a_{\nu-1}\,|\,b_1,\cdots,b_{\mu-1}). \tag{9}$$

The word <u>average</u> here implies a two-fold averaging, viz., first, averaging over all the possible sequences with a fixed position where the sequence $(b_1,\cdots,b_{\mu-1})$ should appear, and second, averaging over all the possible positions of this sequence. The first averaging is mathematically represented by the matrix multiplication in (5), and the second averaging by the summation in (8).

<u>Definition II. If $\lim_{m\to\infty}U^{(m)}(a_1,\cdots,a_{\nu-1}\,|\,b_1,\cdots,b_{\mu-1})$ converges to a unique, non-vanishing limit independent of</u>

$(a_1, \ldots, a_{\nu-1})$, where $(a_1, \ldots, a_{\nu-1})$ can be taken arbitrarily from a certain family of $(\nu - 1)$ - symbol sequences and $(b_1, \ldots, b_{\mu-1})$ can be taken arbitrarily from a certain family of $(\mu - 1)$ - symbol sequences, then we speak of ergodicity with regard to these families.

We shall presently see that the quantity (9) with a fixed initial sequence $(a_1, \ldots, a_{\nu-1})$ and a fixed final sequence $(b_1, \ldots, b_{\mu-1})$ indeed converges to a limit, say:

$$U^{(\infty)}(a_1, \ldots, a_{\nu-1} \mid b_1, \ldots, b_{\mu-1}), \tag{10}$$

but this limit is not necessarily larger than zero, nor is it in general necessarily independent of the initial sequence. In order to understand clearly the situation, let us invoke some well-known mathematical theorems regarding the Markoff chains.[1]

The ordinary Markoff chain formally pertains to a two-symbol correlation probability $(\alpha|R|\beta)$, $(\alpha, \beta = 1, 2, \ldots, M)$:

$$(\alpha|R|\beta) \geq 1 \quad , \quad \sum_{\beta} (\alpha|R|\beta) = 1 . \tag{11}$$

In accordance with the usual rule of matrix multiplication, we further introduce

$$(\alpha|R^m|\beta) = \sum_{\varkappa} \sum_{\lambda} \cdots \sum_{\mu} \underbrace{(\alpha|R|\varkappa)(\varkappa|R|\lambda) \cdots (\mu|R|\beta)}_{m} \tag{12}$$

Then, we have the following theorems:

Theorem II. The quantity defined by

$$U^{(m)}(\alpha|\beta) = \sum_{\ell=1}^{m} \frac{1}{m} (\alpha|R^{\ell}|\beta) \tag{13}$$

1. See for instance W. Feller, Introduction to Probability Theory and its Applications (John Wiley, New York, 1950) p. 307 ff.

for any given pair (α, β) converges to a limit as $m \to \infty$:

$$U^{(\infty)}(\alpha|\beta) = \lim_{m\to\infty} U^{(m)}(\alpha|\beta). \tag{14}$$

Theorem III. The entire set G of symbols ($\alpha = 1,2, \ldots, M$) can be divided into a "vanishing" subset V and a certain number of "closed" subsets $C_i (i = 1,2, \ldots)$ in such a way that

$U^{(\infty)}(\alpha|\beta) = 0$ for α belonging to G, and for β belonging to V,

$U^{(\infty)}(\alpha|\beta) > 0$ for α and β belonging to the same C_i,

$U^{(\infty)}(\alpha|\beta) = 0$ for α and β belonging to different C_i's.

Theorem IV. $U^{(\infty)}(\alpha|\beta)$ is independent of α, if α and β belong to the same C.

Coming back to our original topic, if the correlation-range is two, and if $\mu = \nu$, these theorems can be directly applied to our problem involved in Definition II. If the correlation-range is > 2, we only need to consider a sequence of (ν - 1) symbols collectively as a symbol α. The R's defined in (4) indeed satisfy (11). The cases: $\mu \neq \nu$ can be handled with the help of (6) and (7).

From Theorem II follows quite generally:

Theorem V. The limit (10) exists.

We shall now discuss first the case $\mu = \nu$ in the light of Theorems II, III and IV. According to Theorem III, the entire set of (ν - 1) - symbol sequences is subdivided into a vanishing subset V and a certain number of closed subsets C_i. If the final sequence of (10) belongs to V, then $U^{(\infty)}$ is zero independently of the initial

sequence. For a given final sequence belonging to one of the closed subsets, $U^{(\infty)}$ will be zero if the initial sequence belongs to another closed subset, and will have a constant non-vanishing value insofar as the initial sequence belongs to the same closed subset as the final sequence. Thus:

Theorem VI. When $\mu = \nu$, ergodicity in the sense of Def. II holds if and only if the initial family and the final family are the same closed subset.

In the cases where $\mu > \nu$, we construct an "extended" closed subset D_i of $(\mu - 1)$ symbols by taking those $(\mu - 1)$ - symbol sequences $(b_1, \cdots, b_{\mu-1})$ whose first $(\nu - 1)$ symbols coincide with one of the members of the $(\nu - 1)$ - symbol closed subset C_i and which satisfy the condition:

$$Q(b_1, \cdots, b_{\nu-1} | b_\nu)\, Q(b_2, \cdots, b_\nu | b_{\nu+1}) \cdots Q(b_{\mu-\nu}, \cdots, b_{\mu-2} | b_{\mu-1}) \neq 0 \quad (15)$$

The extended vanishing subset will be composed of all those $(\mu - 1)$ - symbol sequences whose first $(\nu - 1)$ symbols coincide with one of the members of the $(\nu - 1)$ - symbol vanishing subset, or whose first $(\nu - 1)$ symbols coincide with one of the members of some closed subset but whose last $(\mu - \nu)$ symbols violate the condition (15). The entire set of possible $(\mu - 1)$ - symbol sequences are thus covered by the D's and V, and there is no possible overlapping. If the $(\mu - 1)$ - symbol final sequence of (10) is a member of this extended vanishing subset, $U^{(\infty)}$ will certainly vanish whatever the initial sequence may be. If the final sequence belongs to an extended

closed subset D_i, then $U^{(\infty)}$ will vanish for an initial sequence belonging to a C_j different from the one, C_i, which corresponds to D_i, and will have a constant non-vanishing value for any initial sequence belonging to C_i.

Theorem VII. When $\mu > \nu$, ergodicity holds if and only if the initial family is one of the closed subset C_i and the final family is the extended closed subset D_i corresponding to C_i.

In the cases where $\mu < \nu$, we encounter a rather peculiar situation. From a closed subset C_i we construct a retrenched subset E_i of $(\mu - 1)$ - symbol sequences. E_i is the set of those $(\mu - 1)$ - symbols sequences which coincide with the first $(\mu - 1)$ symbols of at least one of the members of C_i. The retrenched vanishing subset is defined as the totality of all those $(\mu - 1)$ - symbol sequences which do not belong to any one of the retrenched closed subsets. In case of the extended closed subsets, a given sequence of $(\mu - 1)$ symbols could not belong to more than one D_i, since the division made in Theorem III does not allow for any overlapping. However, in the present case of retrenched subsets, a given $(\mu - 1)$ -symbol sequence may well belong to more than one E. If the $(\mu - 1)$ - symbol final sequence of (10) belongs to the retrenched vanishing subset, $U^{(\infty)}$ will always vanish. If the $(\mu - 1)$ - symbol final sequence belongs to $E_i, E_j, \ldots, E_k$, then $U^{(\infty)}$ will be zero for an initial sequence belonging to a C different from any one of the corresponding subsets: $C_i, C_j, \ldots, C_k$. For the same final sequence, $U^{(\infty)}$ may thus have different non-vanishing values according as to which one of $C_i, C_j, \ldots, C_k$ the initial sequence belongs.

Theorem VIII. When $\mu < \nu$, ergodicity holds for the initial family identical with one of the closed subset C_i and the final family identical with the corresponding retrenched subset E_i.

In the foregoing considerations, we have systematically omitted the initial sequences belonging to the vanishing subset V. The reason for this is that the $U^{(\infty)}$ depends in this case on the detailed structure of the intersymbol correlation, and that we cannot draw a conclusion of general validity. (Of course, if the final sequence also belongs to V, then $U^{(\infty)}$ vanishes).

Regarding the closed subsets of $(\nu - 1)$ symbols, we should like to mention the following interesting property. We have obviously

$$U^{(\infty)}(a_1, \cdot\cdot, a_{\nu-1} | b_2, \cdot\cdot, b_\nu) = \sum_{b_1} U^{(\infty)}(a_1, \cdot\cdot, a_{\nu-1} | b_1, \cdot\cdot, b_{\nu-1}) Q(b_1, \cdot\cdot, b_{\nu-1} | b_\nu), \quad (16)$$

whence we infer:

Theorem IX. $(b_2, b_3, \cdot\cdot b_\nu)$ is a member of C_i, if there is any symbol b_1 such that $(b_1, b_2, \cdot\cdot b_{\nu-1})$ is a member of C_i and $Q(b_1, b_2, \cdot\cdot, b_{\nu-1} | b_\nu) \neq 0$.

For a given $(b_1, b_2, \cdot\cdot, b_{\nu-1})$ there must be at least one b_ν such that $Q(b_1, b_2, \cdot\cdot, b_{\nu-1} | b_\nu) \neq 0$, on account of (2). Hence:

Theorem X. If $(b_1, b_2, \cdot\cdot, b_{\nu-1})$ is a member of C_i, then there is always a member of C_i whose first $(\nu - 2)$ symbols are $(b_2, \cdot\cdot, b_{\nu-1})$.

Before closing this section, a simple illustration may be given. Suppose the alphabet to be composed of three symbols: S_1, S_2 and S_3,

and to have an intersymbol correlation of range 3:

$$Q(S_1, S_1 \mid S_1) = 1, \qquad Q(S_1, S_2 \mid S_1) = 1,$$
$$Q(S_1, S_3 \mid S_1) = 1, \qquad Q(S_2, S_1 \mid S_2) = 1,$$
$$Q(S_2, S_2 \mid S_2) = 1, \qquad Q(S_2, S_3 \mid S_1) = 1,$$
$$Q(S_3, S_1 \mid S_1) = 1, \qquad Q(S_3, S_2 \mid S_1) = 1,$$
$$Q(S_3, S_3 \mid S_1) = 1.$$

Then the $(\nu - 1)$ symbol subsets are:

C_1 : (S_1, S_1)
C_2 : $(S_1, S_2), (S_2, S_1)$
C_3 : (S_2, S_2)
V : $(S_1, S_3), (S_3, S_1), (S_2, S_3), (S_3, S_2), (S_3, S_3)$

The extended 3-symbol subsets are:

D : (S_1, S_1, S_1)
D : $(S_1, S_2, S_1), (S_2\ S_1\ S_2)$
D : (S_2, S_2, S_2)
V' : all other 3-symbol sequences

The retrenched 1-symbol subsets are:

E : S_1
E : S_1, S_2
E : S_2
V : S_3

We can see the overlapping we have discussed; as a result, $U^{(\infty)}$ with the final sequence (symbol) S_1, for instance, becomes three-valued:

$$U^{(\infty)}(S_1, S_1 \mid S_1) = 1$$
$$U^{(\infty)}(S_1, S_2 \mid S_1) = \tfrac{1}{2}$$
$$U^{(\infty)}(S_2, S_1 \mid S_1) = \tfrac{1}{2}$$
$$U^{(\infty)}(S_2, S_2 \mid S_1) = 0$$
$$\text{All other } U^{(\infty)}(\ \mid S_1) = 1$$

#3. Redundancy

In this section, we shall constantly use a quantity denoted by:

$$P(a_1, a_2, .., a_n) \geq 1. \tag{17}$$

Definition III. The quantity (17) represents the probability, in infinitely long messages, of an arbitrarily taken sequence of symbol-length n being a particular sequence $(a_1, a_2, .., a_n)$.

From this definition follows the normalization condition:

$$\sum_{a_1} .. \sum_{a_n} P(a_1, a_2, .., a_n) = 1 . \tag{18}$$

According to the point of view of the last section, the existence of a unique value of such a probability is not unconditionally guaranteed. Only if the initial sequence $(b_1, . ., b_{\nu-1})$ is limited to within a closed subset, say, C_i, then

$$U^{(\infty)}(b_1, .., b_{\nu-1} | a_1, .., a_n)$$

becomes independent of $(b_1, . ., b_{\nu-1})$, i.e., a function only of $(a_1, . ., a_n)$. If this is the case, we can write

$$U^{(\infty)}(b_1, .., b_{\nu-1} | a_1, .., a_n) = P(a_1, .., a_n) . \tag{19}$$

According to the theorems of the last section, if $(a_1, . ., a_n)$ belongs to C_i, or its extended subset D_i, or its retrenched subset E_i, P will be finite, and otherwise zero. We have therefore to restrict the "infinitely long messages" of Definition III to only those which start with initial sequences belonging to one closed subset. The condition regarding P does not require that all the P's should be non-vanishing, thence the restriction on the final sequences, in the sense of

Definition II, is not necessary. On account of ergodicity, two sequences starting from two different initial sequences of the same closed subset becomes, in the long run, statistically identical. It is true that we can evade the restriction on the initial sequences by giving a certain "weight" to each of the closed subsets, which would lead to a unique value of each P. However, from the point of view that the messages are engendered solely by the correlation probability, this alternative is not acceptable, since it involves an arbitrary "weight" of each closed subset. Our discussion of this section will be based on the assumption that the initial sequences are limited to a single subset. The generalization of the results to the case of "weighted" subsets is very simple.

It should be noted that, as a result of the limitation of the initial sequences to a single subset, it may well happen that some of the generally possible sequences $(a_1, \ldots, a_{\nu-1})$ in the correlation probability $Q(a_1, \ldots, a_{\nu-1} \mid a_\nu)$ actually never happen in the possible messages. Thus the actual range of correlation may become smaller than the range defined with regard to the entire possibilities of the a's. For instance, in the illustration of the last section, if we limit ourselves to the initial subset C_2, all 3-symbol Q's except $Q(S_1, S_2 \mid S_1)=1$ and $Q(S_2, S_1 \mid S_2) = 1$ will become meaningless. These two 3-symbol correlation probabilities reduce to the following two 2-symbol correlation probabilities: $Q(S_1 \mid S_2) = 1$, and $Q(S_2 \mid S_1) = 1$. The range is thus reduced from three to two.

In the empirical point of view, if a population of very long sample messages is given, we can always evaluate (17) by just counting the

frequency of each segment $(a_1, \ldots, a_n)$. However, if we divide this entire population into, say, two groups, the values of (17) may be different in the two groups. This discrepancy may be caused by a difference in correlation probabilities and/or by a difference in the initial sequences. We thus see that the problem of ergodicity is not irrelevant to the empirical point of view. In this section, however, we assume that we have a single population from which the quantities of the type (17) are uniquely determined.

The quantity (17) has, besides (18), the property:

$$\sum_a P(a_1, \ldots, a_k, b_1, \ldots, b_m, a_{k+m+1}, \ldots, a_n) = P(b_1, \cdots, b_m) . \tag{20}$$

This is obvious from the statistical point of view, but can also be verified from the standpoint of (19).

According to (6), we have for $n \geq \nu$

$$P(a_1, \ldots, a_n) = P(a_1, \ldots, a_{\nu-1}) Q(a_1, \ldots, a_{\nu-1} | a_\nu) \cdots Q(a_{n-\nu+1}, \ldots, a_{n-1} | a_n) , \tag{21}$$

or more generally,

$$P(a_1, \ldots, a_n) = P(a_1, \ldots, a_{\mu-1}) Q(a_1, \ldots, a_{\mu-1} | a_\mu) \cdots Q(a_{n-\mu+1}, \ldots, a_{n-1} | a_n) , \tag{22}$$

provided $n \geq \mu \geq \nu$. Equivalence of (21) and (22) can readily be seen with the help of (3) and (6). In particular, for $n = \mu \geq \nu$, we get from (22)

$$Q(a_1, \ldots, a_{\mu-1} | a_\mu) = \frac{P(a_1, \ldots, a_\mu)}{P(a_1, \ldots, a_{\mu-1})} . \tag{23}$$

This is just what should be according to Definitions I and III. (23) may be considered as the definition of $Q(a_1, \ldots, a_{\mu-1} | a_\mu)$ even

for $\mu < \nu$. However, with such Q's with $\mu < \nu$, (22) will not be true, since the Q's with $\mu < \nu$ cannot describe fully the existing correlation.

Substituting (23) into (22), we get

$$P(a_1, \cdot\cdot, a_n) = \frac{P(a_1, \cdot\cdot, a_\mu) P(a_2, \cdot\cdot, P_{\mu+1}) \cdots P(a_{n-\mu+1}, \cdot\cdot, a_n)}{P(a_2, \cdot\cdot, a_\mu) \cdots P(a_{n-\mu+1}, \cdot\cdot, a_{n-1})} , \qquad (24)$$

provided $n > \mu \geq \nu$. The actual range ν is thus the minimum value of μ for which the decomposition (24) is allowed.

For an allowed value of μ, if a further decomposition of range $\mu - 1$ is still allowed, i.e., if $\mu - 1 \geq \nu$, then we get from (24)

$$P(a_1, \cdot\cdot, a_\mu) = \frac{P(a_1, \cdot\cdot, a_{\mu-1}) P(a_2, \cdot\cdot, a_\mu)}{P(a_2, \cdot\cdot, a_{\mu-1})} \qquad (25)$$

for all $(a_1, \cdot\ \cdot, a_\mu)$. But if $\mu - 1 < \nu$, the left side of (25) will not be equal to its right side for at least one sequence $(a_1, \cdot\ \cdot, a_\mu)$. Thus we are led to use (25) as a criterion to determine whether $\mu > \nu$ or not: If (25) holds for all $(a_1, \cdot\ \cdot, a_\mu)$, then $\mu > \nu$; if not, $\mu \leq \nu$. Indeed, if (25) is possible, we have in virtue of (23),

$$Q(a_1, \cdot\cdot, a_{\mu-1} | a_\mu) = \frac{P(a_1, \cdot\cdot, a_\mu)}{P(a_1, \cdots, a_{\mu-1})}$$

$$= \frac{P(a_2, \cdot\cdot, a_\mu)}{P(a_2, \cdot\cdot, a_{\mu-1})} = Q(a_2, \cdot\cdot, a_{\mu-1} | a_\mu) \qquad (26)$$

i.e., Q of range μ is reducible to a Q of range $(\mu - 1)$. In the light of Theorem I, this means that the actual range is $(\mu - 1)$ or less. If (25) breaks down for at least one sequence $(a_1, \cdot\ \cdot, a_\mu)$, then (26) does not hold in general, meaning that the actual range is

larger than $(\mu - 1)$.

Theorem XI. If and only if (25) holds for all $(a_1, ., a_\mu)$, the actual correlation range ν is $(\mu - 1)$ or less.

This criterion is interesting particularly in the empirical point of view, for here the P's, instead of the Q's, are the quantities which are primarily given. The criterion of Theorem XI can be brought to a more concise form by the help of the well-known theorem attributed to W. Gibbs:

Theorem XII. If

$$f_i \geq 0\ ,\ g_i \geq 0\ ,\ \text{and}\quad \sum_i f_i = \sum_i g_i\ ,\ (i = 1, 2, \cdots, r)\ , \tag{27}$$

then

$$W \equiv \sum_i f_i \log f_i - \sum f_i \log g_i \geq 0\ , \tag{28}$$

where the equality holds only when $f_i = g_i$ for all i.

Now, let us call the left-hand side and the right-hand side of (25), respectively

$$f(a_1, \cdots, a_\mu) = P(a_1, .., a_\mu)\ , \tag{29}$$

$$g(a_1, .., a_\mu) = \frac{P(a_1, \cdots, a_{\mu-1})\, P(a_2, .., a_\mu)}{P(a_2, \cdots, a_{\mu-1})}\ , \tag{30}$$

and consider the index i of Theorem XII as a collective index for various possible sequences of symbol-length μ. On account of (18) and (20), the conditions (27) are satisfied, and we obtain

$$\begin{aligned} W_\mu \equiv & \sum P(a_1, .., a_\mu) \log P(a_1, .., a_\mu) \\ & - 2 \sum P(a_1, .., a_{\mu-1}) \log (a_1, .., a_{\mu-1}) \\ & + \sum P(a_1, .., a_{\mu-2}) \log (a_1, .., a_{\mu-2}) \geq 0\ . \end{aligned} \tag{31}$$

Only when (25) holds for all $(a_1, .., a_\mu)$, then $W_\mu = 0$. In other words, for a given value of ν, $W_\mu = 0$ for $\mu > \nu$. this leads to a convenient way to determine the actual range:

Theorem XIII. The actual range ν is the maximum value of μ for which $W_\mu \neq 0$.

The W's defined by (31) will be called "correlation indicies".

For $\mu = 2$, the definition of W_μ in (31) should be understood as meaning

$$W_2 = \sum P(a_1, a_2) \log P(a_1, a_2) - 2 \sum P(a_1) \log P(a_1), \qquad (32)$$

for we have here $q(a_1, a_2) = P(a_1) P(a_2)$.

We shall now proceed to find out the average amount of information carried by a message-segment of length n in a language in which the P's exist. A specific message-segment $(a_1, .., a_n)$ has probability $P(a_1, .., a_n)$. Thus the information per symbol carried by this message-segment is

$$-\frac{1}{n} \log P(a_1, .., a_n).$$

The probability of occurrence of such a message being $P(a_1, .., a_n)$, the average information per symbol for various possible message-segments of length n is given by

$$I_n = -\frac{1}{n} \sum P(a_1, .., a_n) \log P(a_1, .., a_n) \qquad (33)$$

Now, if the existing correlation is of range ν, the P can be decomposed as in (24) with $\mu = \nu$. A straightforward calculation with the help of (18) and (20) gives

$$I_n = I_{n,\nu} \equiv -\frac{1}{n}(n-\nu+1) \sum P(a_1, .., a_\nu) \log P(a_1, .., a_\nu)$$
$$+\frac{1}{n}(n-\nu) \sum P(a_1, .., a_{\nu-1}) \log P(a_1, .., a_{\nu-1}). \qquad (34)$$

For an obvious reason this ν can be the actual minimum range or any ν that is larger than this. Supposing ν in (34) to be the actual minimum range, let us find the error which would be committed by the calculation based on the assumption that the actual range were $\nu - 1$. This is easily found to be

$$I_{n,\nu} - I_{n,\nu-1} = -\frac{(n-\nu+1)}{n} W_\nu \quad . \tag{35}$$

Repeating this process, we obtain

$$I_n - I^\circ = I_{n,\nu} - I^\circ = -\sum_{\mu=2}^{\nu} \frac{n-\mu+1}{n} W_\mu \quad , \tag{36}$$

where

$$I^\circ \equiv I_{n,1} = -\sum P(a_i) \log P(a_i) \quad . \tag{37}$$

Since W_μ vanishes anyway for $\mu > \nu$, we can state:

Theorem XIV. The average information per symbol carried by a message-segment of length n is

$$I_n = I^\circ - \sum_{\mu=2}^{\infty} \frac{n-\mu+1}{n} W_\mu \tag{38}$$

insofar as n is larger than the actual correlation range.

Since the W's are zero or positive, the intersymbol correlation tends to decrease the amount of information. Thus, W_μ can be considered to represent the "strength" of correlation -- strength in the sense of reducing the amount of information. By definition, I_n cannot be negative, thence there is an upper limit to the total "strength" of the correlation:

$$\sum_{\mu=2}^{\infty} \frac{n-\mu+1}{n} W_\mu \leq \sum_{\mu=2}^{\infty} W_\mu \leq I^\circ \quad . \tag{39}$$

For $n \gg \nu$, we obtain from (38),

$$I_n \approx I_\infty = I^\circ - \sum_{\mu=2}^{\infty} W_\mu \qquad (n \gg \nu) \ , \tag{40}$$

showing that if we take a sufficiently long segment as a unit, the information per symbol becomes independent of the length of the segment. This indirectly justifies the usual procedure according to which an infinitely long message is cut into segments of sufficient length and the segments are treated as if they did not have any correlation among them.

The quantity called "redundancy" is defined by[2]

$$R = \frac{I^\circ - I_\infty}{I^\circ} \ . \tag{41}$$

Theorem XV. The redundancy of a language which is characterized by the correlation indices W_μ is given by

$$R = \frac{1}{I^\circ} \sum_{\mu} W_\mu \quad , \quad 0 \le R \le 1 \ . \tag{42}$$

In the illustration of the last section, if we limit the initial sequences to C_2 , we get

$$W_2 = \log 2 \quad , \quad W_3 = W_4 = \cdots = 0$$

$$I^\circ = \log 2 \quad , \quad I_\infty = 0 \quad , \quad R = 100\ \% \ .$$

This last result is not surprising, because the possible infinite sequences are limited to . . . S_1 S_2 S_1 S_2 . . ., which certainly cannot convey any information.

---

2. Stanford Goldman, Information Theory (Prentice-Hall, New York, 1953) p. 45.

TA7
.U6
no.4

29134

Watanabe
A study of ergodicity and redundancy based on intersymbol correlation of finite range.